DOCUMENT RESUME

ED 422 347 TM 028 899

AUTHOR Roberts, Lily; Wilson, Mark
TITLE Evaluating the Effects of an Integrated Assessment System: Changing Teachers' Practices and Improving Student Achievement in Science.
PUB DATE 1998-04-17
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Academic Achievement; Change; *Educational Assessment; Educational Practices; *Evaluation Methods; Hands On Science; *Integrated Activities; Intermediate Grades; Junior High Schools; Middle Schools; Professional Development; Program Evaluation; *Science Instruction; *Scoring
IDENTIFIERS Embedded Items; *Moderation

ABSTRACT

Evaluation findings are presented from a 2-year project to design and field test an embedded assessment system in concert with the development of a middle school science course, "Issues, Evidence, and You," (IEY), by the Science Education for Public Understanding Program (SEPUP) at the Lawrence Hall of Science. IEY is a four-part course that consists of 65 hands-on activities with embedded assessments that draw on related constructions that include the ability to: (1) identify evidence and make tradeoffs; (2) design and conduct experiments; (3) communicate in written and verbal formats; (4) recognize, use, and apply scientific concepts; and (5) work effectively in a group. This paper describes how the effects of this integrated assessment system were evaluated using mixed methods. It presents results for both teacher change in assessment practices and student achievement. Participants were 51 SEPUP teachers and 12 comparisons in non-SEPUP classes. Field test findings indicate that there must be ongoing support to promote assessment reform in the classroom. The use of assessment moderation was the critical component in SEPUP that supported teacher professional development and changed their assessment practices. Assessment moderation brings a group of teachers together to share and score a sample of student work and reach consensus on student scores, thus providing a means for technical quality control in the application of scoring guides. Moderation also gave teachers the opportunity to discuss the instructional applications of student scores and address the conversion of scores to grades to meet reporting requirements. (Contains 2 tables, 2 figures, and 44 references.) (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED 422 347

# Evaluating the Effects of an Integrated Assessment System:

# Changing Teachers' Practices and Improving Student Achievement in Science

Lily Roberts and Mark Wilson

University of California, Berkeley

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL IN
MICROFICHE ONLY HAS BEEN GRANTED BY
Lily Roberts
TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)
2B

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM028899

Paper presented at the annual meeting of the American Educational Research Association, April 17, 1998, San Diego, CA.

Abstract

The purpose of this paper is to describe the evaluation findings from a two-year project to design and field test an embedded assessment system in concert with the development of a middle school science course -- *Issues, Evidence and You* (IEY) -- by the Science Education for Public Understanding Program (SEPUP) at the Lawrence Hall of Science. IEY is a four-part course that consists of 65 hands-on, issue-oriented activities with embedded assessments that draw on related constructs, including: (a) the ability to identify evidence and make tradeoffs (Evidence and Tradeoffs); (b) the ability to design and conduct experiments (Designing and Conducting Experiments); (c) the ability to communicate in written and verbal formats (Communication of Scientific Information); (d) the ability to recognize, use and apply scientific concepts (Understanding Concepts) and (e) the ability to work effectively in a group setting (Group Interaction). These five variables are assessed throughout the year providing teachers with assessments of student progress that can serve to inform instructional practice.

This paper describes how the effects of this integrated assessment system were evaluated using mixed methods. Results for both teacher change in assessment practices and student achievement will be presented. The findings of the field test indicate that there needs to be ongoing support to promote assessment reform in the classroom, simply buying a course with embedded performance tasks is insufficient to promote assessment reform at the classroom level. The use of assessment moderation was the critical component in SEPUP that supported teacher professional development and changed assessment practices. Assessment moderation brings a group of teachers together to share and score a sample of student work, reach consensus on students' scores, and thus it provides a means for technical quality control in the application of scoring guides. Moderation also provides teachers the opportunity to discuss the instructional implications of student scores as well as address the conversion of scores to grades to meet reporting requirements.

The Science Education for Public Understanding Program received funds from the National Science Foundation (NSF) to develop and field-test a year-long, middle school science course with an integrated assessment system, entitled *Issues, Evidence and You (IEY)*. The SEPUP Assessment Project was designed to apply new theories and methodologies in the field of assessment to the practice of teacher-managed, classroom-based assessment of student performance (Wilson & Adams, 1996). In this paper, we describe this integrated assessment system and its function as a medium for teacher change. Further, we summarize the results of the evaluation conducted during the field test of *IEY* in 1994-1995 that highlights the benefits of the integrated assessment system for students and teachers.

## The SEPUP Assessment System

One important feature of the assessment project is that the *assessment system is tied to the important* core concepts and skills that are the goals of a specific curriculum (Sloane, Wilson & Samson, 1996). While compatible with other national, state, or district efforts to implement new forms of student assessment, the SEPUP Assessment System is specially designed for the specific curriculum that SEPUP teachers are using in their classrooms. A second important feature is that assessments are *fully embedded* in the instructional materials and are designed to be an integral part of the instructional process. A third important feature of the system is that the assessment of student performance, and the interpretation and use of information regarding student performance, is *managed by the classroom teacher*. A fourth important feature is that the assessment system is designed to provide teachers with *evidence* of their students' growth, in part, to provide a basis for evaluating their own success.

In congruence with the National Science Education Standards (National Research Council, 1996), the SEPUP *IEY* curriculum is designed to engage students in an "issues-oriented, hands-on" approach to thinking about scientific issues that are relevant to their daily lives (e.g., water, waste, energy, and environment) (SEPUP, 1996). Further, student understanding is assessed in an ongoing manner using the components of the SEPUP Assessment System (Roberts, Wilson, & Draney, 1997). The students' role is changed as they conduct labs and other activities that are designed to help them understand that science is really a way of asking and answering questions and not just a collection of established facts that they are asked to memorize. The teachers' role is that of facilitator in the students' development. The aim is that assessment information becomes a scaffolding mechanism for instructional change that further facilitates student learning. Tobin (1995) indicates that any "critical approach" to teacher development must provide teachers with "autonomy to identify problems in their classrooms and seek solutions that make sense to them" (p. 147). With SEPUP assessment, the long-term goal is for teachers to become autonomous assessors of their students' understanding of science, which will be evidenced by their own professional development in terms of changing instructional and assessment practices.

The eight components of the full[1] SEPUP Assessment System are presented in Figure 1. This assessment system provides a set of tools for teachers to use to: (a) assess student performance on central concepts and skills in the curriculum; (b) set standards of student performance; (c) track student progress over the year on the central concepts; and (d) provide feedback (to themselves, students, administrators, parents, or other audiences) on student progress as well as on the effectiveness of the curriculum materials and classroom instruction. Initially, managing a new classroom-based system of embedded assessment demands much of the teacher. However, we believe that teachers must be recognized as the front-line professionals who will ultimately determine the usefulness or effectiveness of any attempt at educational reform.

---

[1] Note that not all of these components were fully developed at the time of the field test. For example, performance maps have been generated based on empirical evidence collected during the field test, and hence were not available.

Empowering teachers (through the provision of tools, procedures, and support) to collect, interpret and present their own evidence regarding student performance is an important step in the continuing professionalization of teachers in the field of assessment.

---

Insert Figure 1 about here

---

## An Integrated Assessment System as a Medium for Teacher Change: Creating a Community of Judgment

The strategy of employing an integrated assessment system that addresses various elements of high quality professional development evolved from the concept that teachers need to form a *community of judgment* (Wilson, 1994) that can serve as a network much like the "professional-area movement organizations" described by Pennell and Firestone (1996). Wilson (1994) proposed the community of judgment as one approach to changing the "assessment culture" from one that relies primarily on standardized tests to a teacher-centered approach to educational accountability. Efforts to change the assessment culture are direct responses to calls for better assessments given the limitations of standardized tests (Bullough, 1988; Linn, 1987; Office of Technology Assessment, 1992; Shepard, 1989; Spring, 1988; Wiggins, 1989). To change the assessment culture, it seems clear that one must begin with the teacher working directly with the students being measured. Shepard (1989) suggests that teacher enhancement is needed and that teachers should be directly involved in developing and scoring assessments. However, a comprehensive embedded assessment system goes far beyond the scoring task.

### Community of Judgment

Wiggins (1989) suggested that any system of truly authentic assessment should meet several general criteria; the system should be: (a) criterion-referenced; (b) formative; (c) moderated; and (d) clear in the progression of educational development (e.g., score levels on a rubric would reflect the criteria). Wilson's (1994) *community of judgment* for educational accountability satisfies all of these criteria, while adding a fifth essential component -- a substantive *framework* that describes the achievement variables that are valued and thus worth assessing. Broad frameworks already exist in various subject areas, such as the *Benchmarks for Science Literacy* (AAAS, 1993) or the *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989). These broad frameworks serve as guidelines for local development and implementation of curriculum and assessment.

The substantive framework for assessment and accountability suggested by Wilson (1994) would require specific variables within a particular curriculum to be defined and measured, say, throughout a school year to chart students' progress. Tittle, Hecht, and Moore (1993) describe the use of an assessment framework for junior high school mathematics in which they studied students' (a) metacognitive/self-regulatory skills and (b) affective, motivational and attributional beliefs. This assessment framework, however, is not specific to the curricula being taught and hence, less directly useful to the teachers. Wilson's assessment framework is compatible with the constructs embedded in the course, such as the Designing and Conducting Investigations variable in *Issues, Evidence and You* -- thus assessment and instruction are conjoint rather than disparate elements of teaching. In their analysis of the implementation of the assessment framework, Tittle and her colleagues found that it was important to embed the assessment in "activity settings" and that teacher collaboration was critical to ensure that the assessments were meaningful as well as useful in classroom settings (p. 18). Tobin and Gallagher (1987) also found that "activity settings" (i.e., lecture format, whole class interaction, small groups, or individualization) in secondary science classes functioned to "mold the behavior of teachers and students" (p. 62). They found

that high school science teachers in Australia and the United States geared their instruction to the "most able students" in the class (p. 74). Adjusting assessment strategies could counteract this non-equitable mode of teaching.

In addition to a substantive framework, the community of judgment (Wilson, 1994) model also contains (a) a set of *assessment modes* for each variable and scoring guidelines for rating responses; (b) *moderation* of the teachers' judgments that allows for rater improvement (i.e., teacher enhancement in assessment) and rater adjustment (i.e., consensus building to set standards for student performance); and (c) methods of *quality control* that ensure technical measurement standards (e.g., reliability and validity). Assessment modes are a variety of indicators of student achievement, including tasks such as lab reports, performances (e.g., a town meeting) or projects. These modes are designed to promote higher order thinking, to engage students in real-world, meaningful activities, and to be as closely aligned as possible to the criteria of interest. Moderation is, in brief, a scorer-calibration process that can also serve as a mechanism for teacher professional development (Ingvarson, 1990; Linn, 1994; Roberts, Sloane, & Wilson, 1996). Moderation is being used on a very limited basis in this country (Linn & Baker, 1996). Quality control is needed to ensure comparability of teacher judgments across different settings and to authenticate student work (i.e., prevent cheating).

For teachers to assume a central role in this community of judgment for educational accountability, professional development is crucial (Jett & Schafer, 1992; Linn, 1994; Newell, 1992); "teachers need better training and support as they face increasingly complex challenges of classroom assessment" (NCME, 1994). Particular to improving science education, good teaching materials, "vigorous and ongoing" professional development, and new "assessment tools" are basic components of successful reform (Lopez & Tuomi, 1995, p. 78). Richardson (1990) notes that "teachers themselves must be involved in making judgments about what change is worthwhile and significant" and "that practices and ways of thinking outside an individual teacher's own experiences should be introduced into the dialogue" (p. 14). The community of judgment model can be used to engage teachers in making such judgments while benefiting from the interaction with colleagues.

Alternative science assessments will be implemented "only if teachers understand their use and the depth of the content they demand, are empowered to make instructional decisions, and are supported by school districts which encourage teacher change" (Harmon, 1995, p. 46). If teachers are engaged in a community of judgment for educational accountability, then conditions will hold and hence it is far more likely that alternative assessments will be implemented and that they will be meaningful and useful in terms of informing teachers' instructional practices.

The greatest value of creating an assessment system that promotes teacher change is the potential impact it will have for students. Some benefits of revising and revitalizing assessment practices were reported by Roberts (1995) based on a statewide study of the Eisenhower Mathematics and Science Education Grant Programs implemented in California. Roberts noted that changing assessment practices enables teachers to focus on students' work and learning, so that:

- Teachers can diagnose student difficulties and provide appropriate support to scaffold learning and student understanding.
- Teachers can offer a greater variety of assessment activities to support the learning styles of a diversity of students.
- Teachers have evidence that supports their instructional decisions as well as enhances their communication with students, parents, administrators, and others.
- Teachers can fulfill a leadership role in the area of assessment (p. V-2).

Clearly, to meet the challenges of science education reform, the professionalization of teaching demands that teachers change not only their instructional practices, but also their assessment practices. Further, this change in teachers cannot and, in fact, will not occur in a vacuum; therefore providing for a professional community of judgment that serves as a medium for teacher change is essential. Local assessment moderation serves this function in the SEPUP Assessment System.

Local Assessment Moderation

SEPUP teachers engaged in local assessment moderation function as a community of judgment, and this task engagement serves as a medium for high quality professional development. Principles of high quality professional development, such as ongoing support rather than one-shot experiences (Fullan, 1991) and "support for informed dissent" (Little, 1993, p. 138), delineated by these and other researchers (e.g., Sparks & Loucks-Horsley, 1990) were adhered to in the design of local assessment moderation for SEPUP. Local assessment moderation is a critical element of the SEPUP Assessment System as it supports teacher professional growth in the politically charged arena of assessment reform.

An ongoing evaluation of the implementation of the consensus [assessment] moderation process in the state of Victoria in Australia was conducted from 1981, when it was first implemented, to 1984, with a follow-up study in 1989. Based on the four year evaluation in Victoria, Ingvarson (1990, p. 9) noted "the importance of regarding moderation as a complex innovation requiring a considerable period of time for [teachers'] learning and unlearning during its implementation." The Victoria study indicates that the consensus moderation process had "impressive side effects on the professional development and accountability of teachers" (p. 2). Many of the comments from the SEPUP teachers are consistent with the findings of the Victorian evaluation (Ingvarson, 1990). In the Victorian evaluation, it was found that involvement in the consensus moderation process:

(a) added significantly to teachers' skills for assessing student learning;
(b) enhanced teachers' ability to evaluate and improve their teaching;
(c) significantly increased teachers' access to useful ideas for teaching;
(d) enhanced the quality of learning of students;
(e) affected positively participants' teaching in non-project classes; and
(f) supported, rather than intimidated, beginning teachers.

Ingvarson (1990) also reported that the positive responses increased as teachers had more experience with moderation, which again reflects the need for teachers to have time to become knowledgeable and skilled in using this process. This experience factor was also evident during the evaluation of SEPUP (Roberts, 1996).

## Evaluation Design: Integrated Mixed Methods

Chen (1997) suggests that the choice of evaluation design "depends on the contextual circumstances" that surround the particular project to be evaluated (p. 63). Evaluating a complex innovation such as SEPUP Assessment, especially during its developmental stages, lends itself to a mixed methods design. For this study, an integrated mixed-method design was used because the framework for the evaluation is a multilevel model (i.e., teachers nested within SEPUP Centers). This type of mixed-methods design can be differentiated from a component design (e.g., the use of mixed methods to triangulate the results to achieve greater validity) because it achieves a greater integration of the different methods used (Caracelli & Greene, 1997). This study can be characterized as an embedded or nested mixed-methods design, which combines both qualitative and quantitative methods within a multilevel model to explore the affect of this complex innovation

on teachers' assessment, collegial and instructional practices. At the student level, a quasi-experimental pre/post, treatment comparison group design was used (for technical details, see Wilson & Draney, 1997).

The ideal quantitative design would use a multilevel modeling approach. An HLM analysis, for example, can address both multilevel (teachers nested within Centers) and multisite (variation in Centers on the Center-level predictor variables) issues within the same analytical framework. HLM could be used to: (1) examine teacher change on the variables of interest; (2) compare mean differences on teacher change between Centers; and (3) assess within-Center variation attributable to Center-level predictor variables. The first two analyses can also be performed using standard statistical approaches. Given the developmental nature of SEPUP, both the numbers of Centers (or groups) and the numbers of teachers within Centers is too sparse to conduct an HLM analysis. For one solution to address this multilevel, multisite challenge, Roberts (1996) employs an integrated mixed methods design.

The Center-level predictor variables include strength of leadership, level of institutional support, and teacher proximity and collaboration, which were identified through an extensive literature review about context factors that mediate teacher change (Roberts, 1996). These three Center-level predictor variables were hypothesized to affect teacher professional development in terms of changing assessment, collegial and instructional practices.

This integrated mixed-method design also allows for advantages inherent in component designs, including triangulation and complementarity. For example, qualitative information collected from multiple sites can be used to help explain discrepant cases or outliers in the quantitative analyses. According to Seltzer (1992), the cases that generate the most interest are those that are either unusually successful in using an innovation or those that encountered unusual levels of failure or less than anticipated progress. The two methods are complementary because the richness of detail provided by the qualitative information can help to interpret the quantitative evidence, providing information that might otherwise remain beyond the grasp of the evaluation researcher. During the development of the SEPUP curriculum and embedded assessment system, these methods were also used iteratively. For example, qualitative information from interviews and focus groups helped guide the development of the quantitative surveys, then the quantitative results led us to more specific questions to ask when we returned to the Centers to interview teachers and principals, visit classrooms and observe local assessment moderation meetings.

Evaluation Research Questions

The research questions reported here only relate to the individual (i.e., teacher) level of this multilevel, multisite evaluation.

Changing Teachers' Practices.

1.1 Were the changes for the SEPUP teachers greater than those for the non-SEPUP teachers in terms of assessment practices, instructional practices, and collegiality?

1.2 Were the changes for the SEPUP ADC teachers greater than those for the SEPUP PDC teachers in terms of assessment practices, instructional practices, and collegiality?

Student Achievement.

2.1 Were the changes for the SEPUP students greater than those for the non-SEPUP students on the SEPUP variables?

2.2 Were the changes for the SEPUP ADC students greater than those for the SEPUP PDC students on the SEPUP variables?

Subjects

To evaluate SEPUP during the field test, three treatment groups were defined: SEPUP Assessment Development Centers (ADCs); SEPUP Professional Development Centers (PDCs); and a non-SEPUP comparison group. All of the teachers taught middle school science. All SEPUP teachers used the Issues, Evidence and You curriculum and the non-SEPUP teachers used their traditional science curriculum. The SEPUP ADC teachers received the most comprehensive treatment, both the curriculum and the assessment tools designed for SEPUP. The ADC teachers were supported in their use of the SEPUP Assessment System through the use of local assessment moderation meetings.

All teachers had been nominated based on their excellence as science teachers and were willing to participate in the field test. The comparison teachers were the "next on the list" who did not become part of the field test. The distinction between ADC and PDC was at the Center-level rather than the teacher level.

The sample consists of all teachers involved in the 1994-95 field test of SEPUP's IEY and students in at least one of their science classes. There were 26 SEPUP teachers and seven comparison teachers in the ADCs. There were 25 SEPUP teachers and five comparison teachers in the PDCs.

Only the four ADCs that complied with all aspects of the SEPUP Assessment System field test were included in the analyses presented here. Two ADCs were dropped from the analysis because the teachers did not fulfill their obligations with regard to local assessment moderation and the return of evaluation information.

The participating teachers were from various schools and districts within the Centers. Centers were organized a bit differently in each location. Some Centers were located in a single, large school district and a district representative functioned as the ADC director. Other Centers were organized less centrally, for example, one ADC had five SEPUP teachers from five different districts, and one of these teachers also served as the ADC director. In some cases, the Center director was not from the local school or district, but rather a university person involved in science education.

Description of the instrument

Instruments were developed to meet the particular needs of this evaluation research following an exhaustive review of the literature which failed to uncover appropriate instruments. The teacher instrument discussed in this paper is the SEPUP Inventory of Teachers' Assessment, Collegial, and Instructional Practices (SITACIP). The initial SITACIP was pilot tested in 1993-94, then a modified version was used in 1994-95. For the validation study refer to Roberts (1996). Student achievement data was collected using pre-post tests, link tests, and the actual assessment tasks in *IEY* (see Wilson & Draney, 1996, 1997)

The SITACIP survey is a 77-item self-report of teachers' assessment, instructional and collegial practices as well as their perceptions about the usefulness of various assessment strategies. The subsections of the SITACIP fall into three basic question types: (a) frequency of use of various instructional and assessment strategies; (b) Likert-type scales for both collegiality and reasons for assessment strategy choices; and (c) attitudes about the usefulness of different assessment strategies for assessing learning, guiding instruction, and grading.

Table 1 presents the five SITACIP scales that are used in this paper. The scales measure teachers' attitudes, perceptions and practices related to assessment, instruction, and collegiality.

---

INSERT TABLE 1 ABOUT HERE

---

Data Analysis

The statistical techniques used for the teacher data include: (a) t-tests to examine pre to post mean differences within-Centers; and (b) Analysis of Covariance (ANCOVA) using the pretest scale score as the covariate to compare treatment groups (e.g., ADC vs. PDC teachers). The dependent variables were the post-SITACIP scale scores for assessing learning, assessment strategy use, grading, collegiality and instructional activities use. In the multilevel framework, the Center is the independent variable; Center is also a random factor because the sites were selected unsystematically from among all the potential sites to enable generalization of the results of the evaluation.

t-tests were used to examine pre to post mean differences on the SITACIP scales for the three groups (ADC, PDC and comparison teachers). These differences were used to identify the size and direction of change. The acceptable minimum level of significance was set at $p < .05$. The analyses were only done for matched cases (i.e., the pre and post score was available or the teacher was dropped from the analysis).

ANCOVA was used because of limitations imposed by the purposive sampling. There was no possibility under the circumstances to use a block design, therefore a procedure that statistically "blocks" the participants was the most appropriate procedure to use (Keppel, 1991). In this case, the pre-scale scores on the SITACIP were used as the covariates when comparing two groups of teachers. Before making group comparisons, the assumption of homogeneity of slopes or group regression coefficients was tested.

Using ANCOVA, the first comparison made was the difference on the post-scale scores between all SEPUP teachers (i.e., both ADC and PDC) and the comparison teachers. The next comparison made was the difference between the two types of SEPUP teachers (i.e., ADC vs. PDC) on the post-scale scores. The adjusted multiple R squared is presented with the ANCOVA results as the effect size which represents the proportion of variance explained by the group variable.

For technical information about the student achievement analysis see Wilson and Draney (1996, 1997). The student data analysis was done only for overall SEPUP scores rather than for individual SEPUP variables, since there were too few scores on any one variable to be analyzed separately.

## Results

This section presents the evaluation results in terms of the benefits to students and teachers of using the SEPUP Assessment System. The results of the evaluation indicate significant benefits for teachers and students in the Assessment Development Centers.

Benefits to Students

The *IEY* curriculum affects both students' perceptions of the relevance of science (Roberts & Henke, 1997) as well as their science achievement on the SEPUP Variables (Wilson, Sloane, Roberts, & Henke, 1995; Wilson & Draney, 1996, 1997).

*Student Achievement.* As presented in Figure 2, students in the ADCs received the greatest benefits from the course as evidenced by the significant differences between ADC, PDC and comparison students' gains in achievement on the combined SEPUP Variables (Wilson & Draney, 1997). Note that there was steady growth through April, then a slump in ADC scores by the end of the year. However, the end of year average for the ADC students is still significantly larger than that of either the PDC or comparison students. The average gain for the ADC group is roughly three times as great as for either of the other two groups.

The PDC students' average gains on the overall SEPUP Variables were comparable to those of the comparison students who were in traditional science courses. This result indicates that the *IEY* curriculum alone is not sufficient to bring about changes in student achievement on the SEPUP Variables. Remember that the difference between ADC and PDC teachers was the requirement of the ADC teachers to use the components of the SEPUP Assessment System whereas the PDC teachers were only required to teach the *IEY* course. Thus, the effect we observed seems to be in the added value that the assessment system brings to the curriculum.

---

Insert Figure 2 about here

---

The drop at the end of the year among the ADC students is also interesting. There are several possible explanations for this drop in student proficiency at the time of the posttest (Wilson & Draney, 1997). The first is that the posttests were most often administered during the last week or two of the school year, which tends to be a busy time for both teachers and students. Consequently, it is possible that the administration of some of the posttests was somewhat rushed. Also, as can be seen in Figure 2, much more time was spent on the first sections of the curriculum than on the last section. In particular, most ADC classes spent nearly half of the school year on Part 1 of the curriculum, and many spent only a month or so on Part 3. Thus, one would expect that students would have felt much more secure with material from Part 1 than with material from Part 3. Since questions from Part 3 are included in both Link 3 and the Post-anchor test (refer to Figure 2), this may have adversely affected the proficiency estimates of the ADC students at the time of the posttest. These results illustrate how assessment information can be used to provide important formative evaluation information.

*Change in Students' Attitudes.* The IEY curriculum also had a positive effect on students' perceptions of the relevance of science. SEPUP IEY students during the pilot year were significantly more positive about the benefits of science to their daily lives than comparison students (Roberts & Henke, 1997). The analysis of the field test year comparisons between ADC, PDC and comparison students is in progress.

Below are selected comments from students involved in the SEPUP pilot and field tests. The students describe the Assessment System as they experienced it and their perceptions of the range of benefits. The quotes are reprinted directly from student comments.

> SEPUP Assessment is different than the traditional approach in science: *"The form of assessment that we've taken the past 2 years [is] easier. I think that the 0 - 4 grading system is O.K. I guess I like it better than any normal science class. I feel that this way of testing is better because it teaches kids how to explain themselves. Also now we know how*

*to accept both sides of a situation and not just pick the one we like. We consider both sides."*

SEPUP Assessment supports retention of information as it relates to "everyday" life: *"I think that the [assessment] system is much better. I learn more because when you finish a topic you have an [assessment] or project. I learn more if we take it in steps, not just plop down and answer multiple choice questions for an hour and a half. 'Cause soon as we leave so does everything we learn. So I really think it is better 'cause all of the problems are put to everyday things and that makes it easier."*

The scoring and grading system is clear and shows students their progress over time: *"These assessments were different because at the end of each unit, we had to write up what we had learned. I think this was better because it helped me understand the unit more. It also helped me see both sides of an issue. I think the grading system was fair also. A four signifies that we really know what was going on, all the way down to a one which means we didn't understand it at all. I thought that the portfolios were a good idea, because it showed how I progressed throughout the year."*

SEPUP Assessment scaffolds student learning and shows their ability better than a traditional test: *"I understand more when we aren't just reading and taking notes. I think most kids will learn more with the experiments and explanations over reading and taking notes. The portfolio and papers instead of a final exam, for me, was much easier 'cause I do good work, but I choke on tests and exams, so I felt this way showed more about my ability than just an exam."*

SEPUP assessment tasks promote thinking and expression of own ideas: *"I think [SEPUP] assessments are better, because it allows you to think for yourself. You express your own ideas about something. It allows you to come up with an answer for yourself instead of having so many possible answers in a multiple choice test."*

Benefits to Teachers

Teacher change was assessed in terms of assessment, instructional and collegial practices. The evaluation of SEPUP's impact on teacher change examined the differences between ADC and PDC teachers as well as between SEPUP and non-SEPUP (i.e., comparison) teachers (Roberts, 1996). The comparison teachers did not change on any of the measures used. However, the SEPUP teachers did change, and there were interesting differences between the ADC and PDC teachers. Table 2 summarizes the results on measures of teacher change related to their perceptions of the usefulness of alternative assessment strategies for grading and assessing learning as well as their practices in terms of alternative assessment strategy use, instructional strategy use, and collegiality.

---

Insert Table 2 about here

---

*Assessment Practices.* The most interesting findings were for the measures related to assessment practice: teachers' perceptions of the usefulness of alternative assessment strategies for grading and assessing what students have learned about science and use of alternative assessment strategies in their own classrooms. In the beginning, both ADC and PDC teachers perceived alternative assessment strategies to be useful. However, when faced with the reality of using alternative assessments to assess learning, guide subsequent instruction and grade students, ADC teachers turned to the practical issues and concerns that they faced in their classrooms daily, such as scoring large numbers of student papers each week and converting scores to grades. The

moderation meetings were the main support mechanism for teachers as they worked collaboratively to solve these daily dilemmas.

ADC and PDC teachers were significantly different on measures of assessment strategy use and the perceived usefulness of alternative assessment strategies for assessing learning and grading. ADC teachers increased their use of open-ended questions and significantly decreased their use of closed-ended questions. Meanwhile, PDC teachers increased their use of closed-ended questions from the beginning to the end of the school year (i.e., they were increasingly using traditional testing such as multiple choice) and reduced their use of alternative assessment strategies by the end of the field test. This evaluation result indicates that the PDC teachers did not grapple with the same practical and conceptual issues about assessment as their ADC counterparts.

The results clearly underscore the *rhetoric* versus the *reality* in the SEPUP teachers' perceptions of alternative assessment and their classroom practice. The rhetoric of assessment reform indicates that alternative assessment is good, so the PDC teachers provided "socially acceptable" responses and yet continued to use traditional assessment approaches. However, the *reality* of assessment reform is that it is fraught with problems and difficulties that may slow a teacher down in covering the required curriculum, anger parents who want to see their children taught in the same manner that they were, frustrate students who are used to an A through F grading system, and so on. The ADC teachers came face to face with these very real dilemmas of teaching with innovative materials and alternative assessments, so their perceptions of the usefulness of alternative assessments were altered by the reality of implementation. ADC teachers re-assessed the usefulness of alternative assessment strategies through practice and reflection and collegial interaction -- which resulted in small, non-significant decreases rather than anticipated increases on their perceptions of the usefulness of such strategies for assessing learning and grading. However, the ADC teachers remained committed to alternative assessment strategies (Roberts, 1996).

*Instructional Practices.* Both ADC and PDC teachers gained more than half a logit on the instructional activity use scale; the gain was not statistically significant for either group. However, it is substantively important that all SEPUP teachers made some changes in their instruction, which reflect the increased use of student-directed, teacher-facilitated methods and decreased use of traditional methods, such as lecturing to students.

*Collegiality.* Both ADC and PDC teachers gained on the measures of collegiality; the gains were not significant for either group, but the gain was greater for ADC teachers. Looking within the ADC group, collegiality was more evident in ADCs where teachers were all from the same district, where collaboration occurred between moderation meetings, and where there was strong leadership for the group. Strong collegiality enhanced an ADCs' success with local assessment moderation.

*Local Assessment Moderation.* For ADC teachers to manage the SEPUP Assessment System, ongoing support and professional development was necessary for successful implementation. SEPUP *IEY* field test teachers told us that "teachers need to internalize the variables." In other words, teachers need to develop an understanding of what it means, for example, to have their students "using evidence to make tradeoffs." Simply giving teachers *scoring guides* did not sufficiently promote an understanding of this variable (Roberts, 1996). Teachers learned about variables by engaging in local assessment moderation in which they were able to set common standards for judging student performance.

The consensus of the ADC participants was that ongoing support through local assessment moderation was critical to teachers' use of the *IEY* curriculum as well as the SEPUP Assessment System. As teachers used assessment tasks and the scoring guides on their own, they encountered questions or issues of implementation that they shared during moderation meetings with their

colleagues. These meetings provided the ADC teachers the ongoing support necessary to change their instructional and assessment practices.

Below are a few selected comments from ADC teachers related to some of the key roles local assessment moderation played in providing collegial support for learning about the SEPUP Assessment System (Roberts, Sloane, & Wilson, 1996).

> *Building an understanding of embedded assessment and using it effectively.* One teacher noted that although she receives strong collegial support at her school, she feels that moderation with her ADC colleagues was important in helping her understand the Scoring Guides and how to apply them to specific activities. She said that she feels that the moderation process helped her to use the Scoring Guides effectively, with practice making it easier to score each time.
>
> *Feedback on student learning.* One teacher reported that she has been able to realize through moderation that she "hasn't gotten through" to her students about *designing investigations* and that she expects too much of her students sometimes. She indicated that the moderation process allows "time to reflect on what you did right and wrong" and that you have "the opportunity to listen to others and get ideas to use next time."
>
> *Promoting reflective practice.* One teacher noted that she feels that the moderation process has contributed to her professional growth because it makes her examine her teaching and students' learning. She adds, "you reflect on your purpose as a teacher: to see students grow and change, and have students see that they're learning something that's applicable." She reported, for example, that she sees that her students are referring back to prior knowledge and applying it to new tasks.

## Discussion

The National Science Foundation has funded the SEPUP Assessment Project to continue its work by: (a) developing new technologies to enable teachers to use alternative assessments as an integral part of their classroom instruction; (b) developing materials and procedures that enable teachers to take up the new approach to assessment embodied in the assessment system through the use of assessment moderation meetings; (c) designing inservice and technical support procedures necessary to prepare teachers to apply new assessment techniques and to assume effective leadership in classroom-based assessment; and (d) conducting an intensive evaluation of the implementation of a comprehensive classroom-based assessment system and the effects of this implementation on classroom practice, school policies regarding assessment and grading, and teachers' professional development. We are using the evaluation findings reported above and summarized below, which have contributed to our understanding of teacher change and student achievement in the implementation of a classroom-based, embedded assessment system, to guide our current work.

### Student Achievement in Science

Data on student achievement suggest that the systematic use of the assessment system does reinforce and contribute to student performance on the critical SEPUP variables. The link tests and the segments of assessments achieved respectable levels of reliability for purposes of tracking student progress. Students of teachers in ADCs demonstrated gains nearly three times greater on total pre-post scores compared to students of teachers in PDCs or students of a comparison group of teachers (these two groups had approximately equivalent group gains). If these results are generalizable beyond the *IEY* field test, then the results imply that the type of curriculum reform

that *IEY* epitomizes will not succeed in producing dramatic results without a significant and active assessment component. In turn, this will require ongoing support for the professional development of teachers using this assessment approach.

Teacher Professional Development

Teacher change can be described as a *situated process* (Roberts, 1996). Situated in two senses; first, as a process that occurs as a teacher engages in actively integrating theory and practice, and second, situated in terms of the organizational context.

There were few significant changes at the teacher level, but teacher change takes time (Fullan, 1991; Little, 1993; Loucks-Horsley, Hewson, Love, & Stiles, 1998). The results exemplify the *rhetoric versus reality* of assessment reform in middle school science teachers' perceptions and practice. On the one hand, those teachers engaged in using the assessment system reassessed the usefulness of alternative assessment strategies -- through practice, reflection and collegial interaction -- which resulted in decreases (rather than what one might have anticipated to be gains) on several of the teacher change variables. On the other hand, the treatment group of PDC teachers who used only the curriculum and did not engage in local assessment moderation retained rosy perceptions of the usefulness of alternative assessments, but continued to use traditional forms of assessment (e.g., multiple choice tests). The comparison group of non-SEPUP teachers made no significant gains on any of the teacher change variables.

The organizational context factors that affected success with use of the SEPUP Assessment System were identified through a qualitative analysis (Roberts, 1996), and included: leadership; institutional support; and teacher proximity and collaboration. The Assessment Development Centers with strong local leadership, institutional support from the district and/or school level, and teachers who collaborated regularly were the most successful in implementing the SEPUP Assessment System. We concluded that district level support was the most conducive to success. Districts can support teacher professional development in various ways, such as financial resources or release time. Districts are more likely to implement reform activities than an individual school (Levine, 1995) and assessment reform once implemented is more likely to be sustained with district-level commitment (Chrispeel, 1997).

Reference List

American Association for the Advancement of Science (AAAS), Project 2061. (1993). Benchmarks for science literacy. New York: Oxford University Press.

Bullough, Jr., R.V. (1988). Excellence and testing. Ch. 4 in The Forgotten Dream of American Public Education. Ames: Iowa State University Press.

Caracelli, V.J., & Greene, J.C. (1997). Crafting mixed-method evaluation designs. In J.C. Greene & V.J. Caracelli (Eds.), Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms. New Directions for Evaluation, no. 74. San Francisco: Jossey-Bass.

Chen, H-T. (1997). Applying mixed methods under the framework of theory-driven evaluations. In J.C. Greene & V.J. Caracelli (Eds.), Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms. New Directions for Evaluation, no. 74. San Francisco: Jossey-Bass.

Chrispeel, J.H. (1997). Educational policy implementation in a shifting political climate: The California experience. American Educational Research Journal, 34 (3), pp. 453-481.

Fullan, M. (1991). The new meaning of educational change (2nd edition ). New York: Teachers College Press.

Harmon, M. (1995). The changing role of assessment in evaluating science education reform. In R.G. O'Sullivan (Ed.) Emerging roles of evaluation in science education reform (pp. 31-52). New Directions for Program Evaluation, Number 65. San Francisco, CA: Jossey-Bass Publishers.

Ingvarson, L. (1990). Enhancing professional skill and accountability in the assessment of student learning. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. 327 558)

Jett, D.L., & Schafer, W.D. (1992). Classroom teachers move to center stage in the assessment area--ready or not! Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction No. 346 144).

Keppel, G. (1991). Design and analysis: A researcher's handbook (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall Inc.

Levine, H. (1995). The district as the fundamental unit of change. In St. John, M., Levine, H., Håkansson, S., Lopez-Freeman, M., Roberts, L., Shattuck, J., & Von Blum, R., A Study of the California Dwight D. Eisenhower Mathematics and Science Education State Grant Program: Lessons Learned about Designing Professional Development Projects, (Essay III). Inverness, CA Inverness Research Associates.

Linn, R.L. (1987). Accountability: The comparison of educational systems and the quality of test results. Educational Policy, 1 (2), pp. 181-198.

Linn, R.L. (1994). Performance assessment: Policy promises and technical measurement standards. Educational Researcher, 23 (9), 4-14.

Linn, R.L. & Baker, E.L. (1996). Can performance-based student assessments be psychometrically sound? In J.B. Baron & D.P. Wolf (Eds.) Performance-based student assessment: Challenges and possibilities. Chicago, IL: University of Chicago Press.

Little, J.W. (1993). Teacher professional development in a climate of educational reform. Educational Evaluation and Policy Analyses, 15 (2), pp. 129-151.

Lopez, R.E., & Tuomi, J. (1995). Student-centered inquiry. Educational Leadership, 52 (8), pp. 78-79.

Loucks-Horsley, S., Hewson, P.W., Love, N., & Stiles, K.E. (1998). Designing professional development for teachers of science and mathematics. Thousand Oaks, CA: Corwin Press.

National Council on Measurement in Education (NCME). (1994). Conference Highlights Keys to Better Assessment. NCME Quarterly Newsletter, 3 (1). Washington, D.C.: National Council on Measurement in Education.

National Council of Teachers of Mathematics (NCTM). (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: NCTM.

National Research Council. (1996). National science education standards. Washington, D.C.: National Academy Press.

Newell, S.T. (1992). Science teachers' perspectives on alternate assessment. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1992.

Office of Technology Assessment. (1992). Testing in American Schools: Asking the Right Questions. Washington, D.C.: Congress of the United States.

Pennell, J.R., & Firestone, W.A. (1996). Changing classroom practices through teacher networks: Matching program features with teacher characteristics. Paper presented at the annual meeting of the American Educational Research Association, New York, NY, April.

Richardson, V. (1990). Significant and worthwhile change in teaching practice. Educational Researcher, 19 (7), pp. 10-18.

Roberts, L. (1995). Assessment as a vehicle for teacher professional development: Moving teachers further along the continuum. In St. John, M., Levine, H., Håkansson, S., Lopez-Freeman, M., Roberts, L., Shattuck, J., & Von Blum, R., A Study of the California Dwight D. Eisenhower Mathematics and Science Education State Grant Program: Lessons Learned about Designing Professional Development Projects, (Essay V). Inverness, CA Inverness Research Associates.

Roberts, L.L.C. (1996). Methods of evaluation for a complex treatment and its effects on teacher professional development: A case study of the Science Education for Public Understanding Program. Unpublished dissertation, University of California at Berkeley.

Roberts, L., & Henke, R. (1997). Mapping middle school students' perceptions of the relevance of science. In Wilson, M., Engelhard, G., & Draney, K. (Eds.), Objective Measurement: Theory into Practice (Vol. 4). Norwood, NJ: Ablex.

Roberts, L., Sloane, K., & Wilson, M. (1996). Local assessment moderation in SEPUP. Paper presented at the Annual Meeting of the American Educational Research Association. New York, April.

Roberts, L., Wilson, M., & Draney, K. (1997). The SEPUP Assessment System: An overview. University of California, Berkeley: BEAR Report Series, SA-97-1.

Seltzer, M.H. (1992, November). Linking the quantitative and qualitative traditions in multi-site evaluations: A multi-level modeling approach. Paper presented at the American Evaluation Association conference, Seattle, WA. [Draft cited with permission from author].

SEPUP. (1996). SEPUP: Science Education for Public Understanding Program (pamphlet). Berkeley, CA: SEPUP, Lawrence Hall of Science.

Shepard, L.A. (1989). Why we need better assessments. Educational Leadership, 46 (7), p. 4-9.

Sloane, K., Wilson, M. & Samson, S. (1996). Designing an embedded assessment system: From principles to practice. Paper presented at the Annual Meeting of the American Educational Research Association. New York, April.

Sparks, D., & Loucks-Horsley, S. (1990). Models of staff development. In Houston, W.R. (ed.) Handbook of Research on Teacher Education. New York: Macmillan Publishing.

Spring, J. (1988). Conflict of interests: The politics of American education. New York: Longman.

Tittle, C.K., Hecht, D., & Moore, P. (1993). Assessment theory and research for classrooms: From taxonomies to constructing meaning in context. Educational Measurement: Issues and Practice, 12 (4), pp. 13-19.

Tobin, K. (1995). Teacher change and the assessment of teacher performance. In B.J. Fraser & H.J. Walberg (Eds.), Improving Science Education. Chicago, IL: University of Chicago Press.

Tobin, K., & Gallagher, J.J. (1987). The role of target students in the science classroom. Journal of Research in Science Teaching, 24 (1), pp. 61-75.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, pp. 703-713.

Wilson, M. (1994). Community of judgment: A teacher-centered approach to educational accountability. In Office of Technology Assessment (Ed.), Issues in Educational Accountability. Washington, D.C.: Office of Technology Assessment, United States Congress.

Wilson, M., & Adams, R.J. (1996). Evaluating progress with alternative assessments: A model for Chapter 1. In M.B. Kane (Ed.), Implementing performance assessment: Promise, Problems and Challenges. Hillsdale, NJ: Lawrence Erlbaum Publishers.

Wilson, M., & Draney, K. (1996). Mapping student progress with embedded assessments. Paper presented at the Annual Meeting of the American Educational Research Association. New York, April.

Wilson, M., & Draney, K. (1997). Developing maps for student progress in the SEPUP assessment system. Paper presented at the Middle School meeting of American Association for the Advancement of Science (AAAS). Seattle, WA, February.

Wilson, M., Sloane, K., Roberts, L., & Henke, R. (1995). SEPUP Course 1, Issues, Evidence and You: Achievement evidence from the pilot implementation. Berkeley, CA: Graduate School of Education, University of California.

Table 1

Descriptions of SITACIP Scales

| Scales | Type of Response Categories | Range of Responses |
|---|---|---|
| Assessing Learning - Perception of usefulness of different assessment strategies for developing an understanding of what students know. | Likert-type attitude scale (useless, somewhat useful, quite useful, very useful) | 1 - 4 |
| Assessment Strategy Use - Use of different assessment or evaluation strategies (e.g., lab reports; open-ended questions). | Frequency teacher uses each assessment or evaluation strategy (never to more than once per week) | 1 - 5 |
| Collegiality - Time spent discussing different topics with other teachers (e.g., assessment/ evaluation strategies). | Frequency teacher discusses topics with other teachers (often, i.e., never to more than once per month) | 1 - 4 |
| Grading - Perception of usefulness of different assessment strategies for grading. | Likert-type attitude scale (useless, somewhat useful, quite useful, very useful) | 1-4 |
| Instructional Activity Use - Use of different instructional strategies to teach middle school science (e.g., "You ask students questions about the material they've read or heard about.") | Frequency teacher uses each instructional strategy listed (never to every day) | 1 - 5 |

Table 2

Summary of t-tests and ANCOVA Results for Selected SITACIP Scales

| Scale | t-test | | | | | | ANCOVA | | |
|---|---|---|---|---|---|---|---|---|---|
| | ADC | | | PDC | | | ADC vs. PDC | | |
| | n | Post[a] | M Diff.[b] | n | Post | M Diff. | df[c] | F[d] | Adj. $R^2$ |
| Grading | 10 | 0.52 | -.96* | 12 | 1.46 | .82* | 19 | 9.39** | .49 |
| Assessment Strategy Use | 19 | 0.68 | .20 | 11 | .04 | -.34 | 27 | 8.30** | .31 |
| Assessing Learning | 10 | 1.00 | -.22 | 12 | 1.40 | .47* | 19 | 3.34 | .18 |
| Instructional Activity Use | 10 | 1.02 | .58 | 13 | 1.24 | .53 | 20 | .18 | .01 |
| Collegiality | 18 | 1.87 | .33 | 11 | 1.62 | .21 | 26 | .11 | .00 |

[a] Post: Post-test scale score in logits.

[b] M Diff.: Mean difference pre to post in logits.

[c] Degrees of Freedom associated with the S within group error.

[d] F ratio.

* $p < .05$.

** $p < .01$.

Figure 1. SEPUP Assessment Components at a Glance

SEPUP Variables - Five variables which represent student learning in terms of the core concepts of IEY.

Assessment Tasks - Activities that are an integral part of regular instruction.

Scoring Guides - Rubrics that establish baseline criteria for assessing levels of student performance.

Assessment Blueprints - Sequential list of IEY activities and opportune points in instruction for assessing student learning.

Exemplars - Examples of student work that have been scored and moderated for each variable and score level.

Assessment Moderation - A process by which teachers discuss and reach consensus on local standards for scoring student work.

Performance Maps - Graphical representations of student development on the SEPUP Variables.

Link Tests - Additional assessment activities for teachers' use at major course transitions that are also based on the SEPUP Variables.

Figure 2. Gains for the ADC, PDC, and comparison groups on the combined SEPUP variable during the 1994-95 school year